HW4

學號： R06323011　系級：經濟碩一　姓名：葉政維

A. PCA of colored faces

A.1. (.5%) 請畫出所有臉的平均。（左下）

A.2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。（右下 4 小圖）

A.3. (.5%) 請從數據集中挑出任意四個圖片，並用前四大 Eigenfaces 進行 reconstruction，並畫出結果。（左下下 4 小圖為原圖（取自圖檔 0, 4,8,12.jpg），而右下下則為對應的重建圖）

A.4. (.5%) 請寫出前四大 Eigenfaces 各自所佔的比重，請用百分比表示並四捨五入到小數點後一位。（前四 Singular Value 比重為：4.1, 2.9, 2.4, 2.2；Eigen 則為 21.6, 10.9, 7.2, 6.1）

B. Image clustering

B.1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

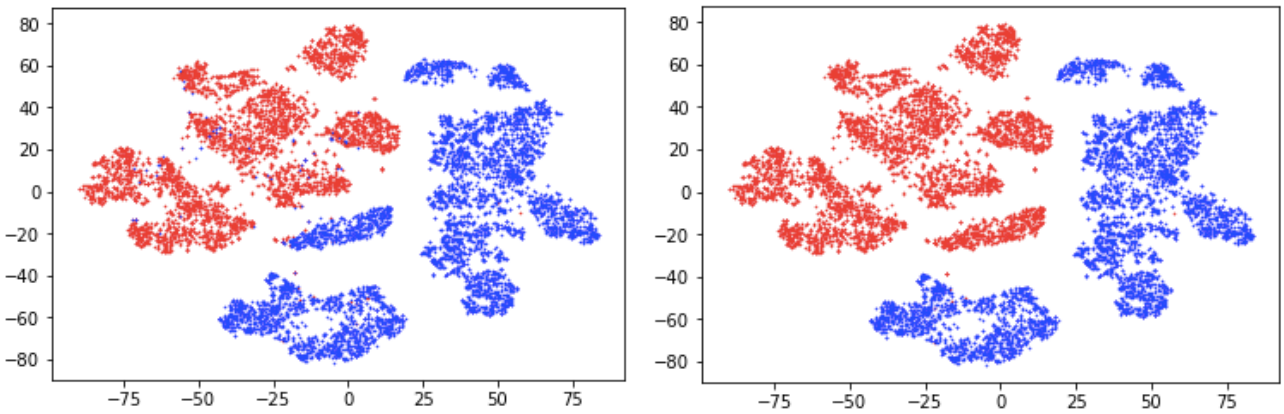
（collaborator：r05323040 田家駿，此外 CAE 參考自 Keras Blog）

本題使用兩種方法：（1）Convolutional Auto Encoder + PCA + Kmeans （2）Convolutional Auto Encoder + Kmeans。我在 PCA 上沒試出太好的結果，所以就不列出來了。CAE 最終使 784 維的資料降至 128 維，而再次使用 PCA 降至 24 維，最終使用 Kmeans 分兩群。此外，圖片都有經過去除雜訊的前處理，具體而言，我將 pixel 值為 255 或 0 的值由周遭的 pixel median 取代。

最終（1）、（2）的 private F1 score 分別為：0.66351、0.99967，儘管使用 Autoencoder 有機會得出一些 pixel 上非線性組合出的特徵，再經過 PCA 降維以去除一些雜訊，可以幫助 Kmeans 更聚焦於那些真的有解釋力的 features (及其線性組合) 身上。

B.2. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。

B.3. (.5%) visualization.npy 中前 5000 個 images 跟後 5000 個 images 來自不同 dataset。請根據這個資訊，在二維平面上視覺化 label 的分佈，接著比較和自己預測的 label 之間有何不同。



依照 B1 的流程將此份資料先 encode 至 128 維，在用先前的 PCA 降至 24 維，最後再用 Kmeans 做預測 label。為進行比較，刻意選了一個不好的 Kmeans 結果（大部分的結果都是完全分對）。

左上為預測結果視覺化，右上則是依據真實標記，其中紅色來自數字資料集，藍色來自服飾資料集。比較發現，大約在中心有一群樣本被分錯了。這群被分錯的幾乎都是數字「1」，這可能是因為這群人和某些連身裙/大衣看起來形狀、輪廓很類似（就像胖一點的 1）。

C. Ensemble learning

C.1. (1.5%) 請在 hw1/hw2/hw3 的 task 上擇一實作 ensemble learning，請比較其與未使用 ensemble method 的模型在 public/private score 的表現並詳細說明你實作的方法。（所有跟 ensemble learning 有關的方法都可以，不需要像 hw3 的要求硬塞到同一個 model 中）

本題嘗試在 hw1 上進行 bagging，preprocessing 如同 hw1 的 report 並使用 18*9 + 1 個 features。比較方式是利用 validating set 的 RMSE，由於 hw1 的資料順序有意義，因此有先經過 shuffle 再抽取約 10%的資料進行 validating。

Bagging 的作法為：在原始的 train set 上 resample 並進行訓練，依 bagging 次數可以得到對應的預測值，最終將這些預測值取平均。未 bagging 和 bagging 的結果依序為（不同的 validating set）：5.2567、5.2591（5.5955、5.5931；4.9996、5.004）。結果顯示 Bagging 並沒有帶來太多幫助（假設我沒做錯），一個可能的原因是使用的模型並沒有複雜到需要透過 Bagging 來降低 Variance。